A Implementation Details

A.1 Model Architecture and Training Schedule

We initialize the canonical 3D Gaussians using 10,000 uniformly sampled point cloud on the robot mesh, and train the canonical Gaussians for 2,000 steps using the L1 image reconstruction loss. Simultaneously, we train our implicit LRS model using the chamfer distance between canonical and posed robot point clouds obtained from simulation. Finally, we train the canonical gaussians, the LRS model, and our appearance deformation network jointly on the L1 image reconstruction loss until convergence. The overall process takes 15-30 minutes on a single NVIDIA 3090 GPU depending on the robot embodiment.

Implicit LRS and appearance deformation modules of our method are modeled using 4-layer 256 hidden dimension lightweight MLPs. We encode coordinates using Fourier features as inputs to the networks.

We keep all canonical 3D Gaussian training hyperparameters consistent with the original 3D Gaussians codebase, and we use the same training hyperparameters across all robot embodiments.

A.2 Text to Robot Pose with CLIP

In section 3.3.1 as our CLIP model, we use `openai/clip-vit-base-patch32` from Huggingface. We minimize the dot product between language and image embeddings that are output by their respective towers.

A.3 Text to Action Sequences with Video Model

To obtain the generative video model shown in the paper, we fine-tune SVD-XT, which can generate 24-frame videos from an image. To enable language conditioning of this model, we replace the OpenCLIP `laion2b_s32b_b79k` model image embeddings with text embeddings from the same model. With this modification, we fine-tune this model on 256x256 videos from the OpenEmbodiment ASU dataset at 10fps following the implementation of [15] until convergence, which takes 12 hours on 2 A100 GPUs. The initial image conditioning and the prompt for the generated videos in the paper are chosen from a set of 12 validation episodes not seen during training.